I. ПРОБЛЕМЫ КОРПУСНОЙ ЛИНГВИСТИКИ В СОПОСТАВИТЕЛЬНЫХ ИССЛЕДОВАНИЯХ

УДК 81’42(045)

Богоявленская Ю. В.

Уральский государственный педагогический университет, Екатеринбург, Россия

СОПОСТАВИТЕЛЬНЫЙ ОБЪЕКТНО-ОРИЕНТИРОВАННЫЙ КОРПУС: ОПРЕДЕЛЕНИЕ ПОНЯТИЯ И ПРИНЦИПЫ ФОРМИРОВАНИЯ

В настоящей статье рассмотрены проблемы обращения к размещенным в свободном доступе корпусам для проведения исследований в русле сопоставительной корпусной лингвистики и предложено их решение. Корпусы разных языков, представленные в сети Интернет, существенно различаются по целям и принципам создания, объему, структуре, содержанию, типам разметки, хронологическим рамкам и другим характеристикам, что не может обеспечить объективность полученных при их анализе данных в сопоставительном исследовании. На первом этапе исследования мы проводим анализ работ, в которых обсуждается наполнение и принципы формирования сопоставительного корпуса, уточняем это понятие и определяем место сопоставительного корпуса в типологии многоязычных корпусов. На следующем этапе, ставя перед собой цель создания собственного сопоставительного корпуса для исследования парцелляции как когнитивно-дискурсивного феномена, мы вводим понятие сопоставительного объектно-ориентированного корпуса, понимая под ним специализированный двуязычный или многоязычный корпус, предназначенный для изучения конкретного лингвистического объекта. Далее формулируются принципы создания корпуса с учетом его специфики: принципы репрезентативности, сбалансированности и сопоставимости. Последний принцип опирается на соблюдение ряда квантитативных и квалитативных критериев сборки сопоставительного корпуса. Только последовательная реализация данных принципов позволит обеспечить достоверность получаемых в ходе корпусного анализа данных об исследуемом объекте. Для сборки сопоставительного объектно-ориентированного корпуса и его обработки используется компьютерная программа *Linguistica*, специально разработанная для решения задач исследования.

Ключевые слова: (сопоставительная) корпусная лингвистика, сопоставительный корпус, сопоставительный объектно-ориентированный корпус, принципы формирования корпуса.

Сведения об авторе: **Богоявленская Юлия Валерьевна**, доктор филол. наук, доцент кафедры романских языков, Уральский государственный педагогический университет (г. Екатеринбург); e-mail: jvbog@yandex.ru, ORCID ID 0000-0002-4500-1491.

Введение

Корпусная лингвистика является активно развивающимся направлением, которое занимается разработкой, созданием и использованием текстовых корпусов. На данный момент существует уже большое количество корпусов на разных языках, совершенствуется технология их формирования, разметки и вывода статистических данных. Однако применение корпусов в сопоставительных исследованиях до сих пор вызывает ряд проблем и ограничений. Актуальность настоящей работы, таким образом, определяется необходимостью анализа сложившейся ситуации и поиска решения, которое смогло бы

удовлетворить современного исследователя, работающего в русле сопоставительной корпусной лингвистики. Актуальность работы также обусловлена необходимостью уточнения понятийно-терминологического аппарата исследования. Например, ключевое понятие сопоставительной корпусной лингвистики «сопоставительный корпус» обсуждается преимущественно в зарубежной лингвистике, имеет противоречивые дефиниции и нуждается в уточнении. Не определено место сопоставительных корпусов в типологии многоязычных корпусов. В работе также уточняются введенные нами ранее понятия «объектно-ориентированный корпус» (далее ООК) и «сопоставительный объектно-ориентированный корпус» (далее СООК), которые выступают объектом нашего исследования.

Предметом исследования являются принципы формирования данного типа корпуса. Практическая значимость работы видится нам в разработке требований к сборке СООК. Практическая ценность работы также связывается нами с возможностью применения ее результатов в преподавании курсов «Сопоставительная лингвистика», «Сравнительная типология языков», «Корпусная лингвистика», «Компьютерная лингвистика» и др. В качестве материала исследования использован сопоставительный французско-русский корпус газетных статей, созданный нами при помощи программы *Linguistica* (Свидетельство о государственной регистрации № 2014660349 в Роспатенте от 06.10.2014).

1. Проблемы использования современных корпусов в сопоставительных исследованиях

В настоящее время существует значительное количество корпусов, представляющих собой массивы текстовых данных на разных языках, однако возможность их использования исследователем ограничена из-за ряда серьезных проблем:

1.1. Доступные в сети Интернет *корпусы ориентированы на анализ лексических или грамматических явлений*, а единицы коммуникации, не имеющие стандартных способов выражения, дискурсивные явления, прагматические особенности текстов, речевых актов и т. д. в них не размечаются, поскольку создание подобной автоматической разметки, по крайней мере, на данном этапе развития компьютерной лингвистики, невозможно. В связи с этим некоторые исследователи отрицают пользу корпусов в изучении текста и дискурса [1].

1.2. Следующая проблема – это *ограниченный доступ или отсутствие доступа к некоторым корпусам*. В частности крупнейший корпус, включающий произведения французской литературы с XX в. по XXI в. *Frantext* доступен только для ограниченного онлайн-поиска, требуется подписка от имени академического / образовательного учреждения. Многие корпусы английского языка являются платными. Полный доступ к Британскому, Американскому национальным корпусам, Международному корпусу английского языка и некоторым другим корпусам в режиме онлайн отсутствует. Но есть и ряд доступных корпусов – это корпусы, созданные Марком Дэвисом и корпусы, доступные с сайта университета Лидс.

1.3. Имеющиеся корпусы *не всегда позволяют получить контекст употребления и доступ к полному тексту произведения*, что снижает возможность адекватной интерпретации исследуемого явления в тексте. Чаще всего исследователь получает доступ к одной строке, одному предложению или некоторому количеству символов (ограничение зависит от корпуса), в котором использовано запрашиваемое слово или выражение.

1.4. Основная проблема использования корпусов в сопоставительных исследованиях – *разный объем, структура корпусных массивов, время создания и типы текстовых документов*. В частности при сопоставлении французских, английских и русских корпусов обнаруживаются все вышеперечисленные (1–3) проблемы. Для большей наглядности сравним доступные корпусы этих языков. *Французский корпус Лейпцигского университета* [2] представляет собой базу данных, включающую 700 млн. слов. Он содержит тексты французской прессы (около 350 млн. слов), веб-страницы и материалы Википедии. Данный корпус позволяет получить список примеров размером в одно предложение на слово-запрос с указанием ссылки на источник, но без даты, а также список правосторонних и левосторонних

коллокатов, т. е. слов, которые часто используются с запрашиваемым словом. Корпус *Frantext* [3], созданный Университетом Нанси, содержит размеченные тексты французских литературных произведений X–XXI вв. в объеме 286 млн. слов и имеет ограниченный доступ. В открытом доступе находится канадский корпус французского языка *Lexiqum* [4] (около 229 млн. слов), созданный Монреальским Университетом для изучения совместной лексической встречаемости и синтаксической сочетаемости различных выражений. Количество примеров на запрос ограничено 500 предложениями длиной не более 200 символов. В корпусе используется только метатекстовая разметка. На данный момент не существует национального корпуса французского языка, сопоставимого по структуре, разметке, функциям, например, с *Британским Национальным корпусом* [5] (100 млн. слов)**,** *Корпусом Современного Американского английского языка COCA* [6] (445 млн. слов) или *Национальным корпусом русского языка* [7] (более 500 млн. слов). НКРЯ – сбалансированный корпус с шестью видами разметки. Он включает произведения художественной литературы, научные тексты и публицистику с сер. XVIII до нач. XXI в. Имеется также *Компьютерный корпус газетных текстов русского языка конца XX века* [8] (1994–1997 гг.) с метатекстовой, морфологической, синтаксической, семантической разметкой.

Как видим, представленные корпусы не являются эквивалентными по объему, содержанию, структуре, хронологическим рамкам, возможностям доступа и т. д., что препятствует применению данных корпусов в сопоставительных исследованиях, поскольку результаты, полученные при анализе данных, отражают разные текстовые массивы, собранные по различным принципам и критериям. Возможен ли выход из этого положения? Как покажет наше дальнейшее исследование, ответ на этот вопрос положительный, но перед тем, как ответить на него, обратимся к уточнению понятия «сопоставительный корпус», представим типологические характеристики собранного корпуса и введем понятие «сопоставительный объектно-ориентированный корпус».

2. Уточнение понятийно-терминологического аппарата исследования

2.1. Понятие «сопоставительный корпус»

Как известно, корпусы могут быть одноязычные и многоязычные. Одним из видов многоязычных корпусов являются *сопоставительные корпусы* (фр. *corpus comparables*; англ. *comparable corpora*), которые, к сожалению, остаются до сих пор вне интересов российской корпусной лингвистики. Зарубежные исследователи отмечают высокую значимость сопоставительных корпусов текстов и ведут активные исследования, связанные с принципами их построения и использования [9–14] и др.

Обзор зарубежных работ выявил существенные различия в определениях понятия «сопоставительный корпус», который понимается либо как двуязычный (мультиязычный), либо как моноязычный или смешанный языковой ресурс, либо ассоциируется с параллельными корпусами.

Л. Боукер и Дж. Пирсон предлагают рассматривать сопоставительные корпусы как «несколько аналогичных по структуре одноязычных подкорпусов, имеющих сходную структуру, не включающих переводные тексты и обладающих сходными характеристиками» (*здесь и далее перевод наш, Ю. Б.*) [15, с. 93]. Авторы уточняют, что в сопоставительный корпус можно включать текстовые документы только определенного типа, так как подобный корпус «фокусируется на особых аспектах языка. Он может быть ограничен рамками языка для специальных целей какой-либо предметной области, специфическим типом текстов, особой разновидностью языка или языком, используемым членами какой-либо демографической группы (например, тинейджеров)» [15, с. 12].

Ряд авторов относят к сопоставительным корпусам иллюстративный материал на двух и более языках, не содержащий переводных текстов, создаваемый с целью сопоставления языков [16, с. 69].

Э. Афли, Л. Барро и О. Швенк добавляют критерий близкой тематики и содержания текстов: «Сопоставительный корпус – это корпус текстов на двух разных языках,

не содержащий параллельных текстов в строгом смысле этого термина, но которые содержат аналогичную информацию» [10, с. 448].

Однако существует и другая точка зрения, согласно которой для сопоставительного корпуса допустимо включение переведенных текстов. В частности, С. Роф понимает под сопоставительным корпусом коллекцию данных (текстов) на нескольких языках, отобранных независимо друг от друга, но которые часто содержат части, являющиеся переводными [17]. А. Мак Энери, З. Ксиао акцентируют внимание на том, что подобные корпусы используются в различных целях (сопоставительные исследования и исследования перевода) и строятся в соответствии с различными принципами: сопоставительные подкорпусы должны быть примерно одинакового объема, представлять одни и те же жанры, предметные области и хронологические периоды. Для параллельных корпусов эти требования нерелевантны [12, с. 3].

Целый ряд исследователей трактуют сопоставительный корпус как моноязыковой, построенный на нескольких подкорпусах [18, 19]. Однако и здесь нет единодушия. Для Ф. Занеттин, О. Куло и др. один из корпусов должен включать письменные тексты на одном языке, а другой – тексты, переведенные на этот язык, что необходимо для изучения трансформаций, возникающих при переводе [20, 21]. Б. Картони и Л. Дележе подразумевают под сопоставительным корпусом объединение нескольких корпусов текстов одного языка, представляющих разные типы речи и разные тематики [11]. М. Гидер, как и некоторые другие исследователи, считает, что понятие сопоставительного корпуса распространяется исключительно на переводные (выровненные) корпусы [22, с. 95]. Ф. Скарпа расширяет границы моноязычного или многоязычного сопоставительного корпуса и строит свои исследования как на параллельных (оригиналы на языке А и их переводы на языке В), так и на непереводных текстах, объединенных близкой тематикой: «сопоставительные корпусы – это корпусы, состоящие из переводов и непереводных оригинальных текстов» [23, с. 124]. С нашей точки зрения, данный вид корпусов логично и оправданно использовать с теми целями, на которые указывает автор, но в исследованиях, ориентированных на сопоставление языков, они вряд ли приемлемы. Как справедливо замечает Л. Герио, «... источники, как правило, сильно влияют на перевод. На самом деле различные обороты и лексика переводного текста обнаруживают тесную связь с текстом-источником» [9, с. 13].

Мы считаем, что не совсем правомерно объединять столь разные по содержанию и целям корпусы под понятием «сопоставительный корпус» и предлагаем следующую типологию, позволяющую провести их более точное разграничение по принципу количества языков и учитывающую сферы их применения и функции:

1) **одноязычные сравнительно-сопоставительные корпусы**, охватывающие корпусы оригинальных и непараллельных переводных текстов на одном языке; диахронические корпусы, создаваемые с целью сопоставления языка, например, XVII и XXI вв., корпусы национальных вариантов одного языка и др., обслуживающие исследования языковых явлений, происходящих в рамках одного языка, и некоторых аспектов перевода;
2) **многоязычные корпусы**:
   - *параллельные (выровненные) корпусы*, т. е. корпусы, включающие тексты, содержащие предложения на одном языке и соответствующие им предложения на втором, третьем и т. д., создаваемые для изучения различных аспектов перевода, а также для обучения методам и приемам перевода;
   - *сопоставительные корпусы*, содержащие текстовые массивы на двух и более языках, относящиеся к одному (или более, в зависимости от поставленных задач) типу речи, функциональному стилю, дискурсу и т. д., сфера применения которых ограничивается сопоставительными исследованиями с возможностью дальнейшего применения результатов в практике преподавания языков; в комбинации с параллельными корпусами могут использоваться для исследований перевода;

– *комбинированные корпусы*, включающие моноязычные подкорпусы и подкорпусы текстов на разных языках, объединяющие перечисленные выше характеристики. Структура, содержание и область применения подобных корпусов зависит от комплекса задач, которые ставит перед собой исследователь.

Таким образом, мы относим сопоставительные корпусы к многоязычным корпусам и отграничиваем их от параллельных и комбинированных корпусов, выполняющих иные функции. Следует отметить, что сопоставительные корпусы обладают целым рядом преимуществ. В первую очередь, это преимущества практического порядка, так как корпусы этого типа в отличие от параллельных корпусов имеют более высокое качество, которое не зависит от субъективного фактора – личности переводчика и качества перевода. Во-вторых, как замечает П. Фунг, материал для них более доступен [24]. В-третьих, сопоставительные корпусы имеют важное преимущество: они не связаны друг с другом, тексты одного корпуса не являются переводными, а потому «использование сопоставительного корпуса позволяет получить непосредственный доступ к реальному употреблению слов в каждом языке и, следовательно, избежать смещений, вызываемых переводом» [25, с. 3].

2.2. Типологические характеристики сформированного сопоставительного корпуса и понятие «сопоставительный объектно-ориентированный корпус»

Обзор корпусов, представленный в п. 1, показывает, что воспользоваться ими для проведения сопоставительного исследования, в частности, французского и русского языков довольно затруднительно. Как отмечалось, корпусы этих языков существенно различаются по объему, принципам построения, разметке, хронологическим рамкам текстов, составу и т. д. Второй причиной, ограничивающей обращение к имеющимся корпусам, стал тот факт, что объектом нашего исследования является парцелляция – явление, которое не имеет лексических средств выражения или формально-грамматических показателей. Работа с имеющимися корпусами строится на лексических запросах (слово, лемма, выражение), вводимых в поисковую строку. Задать поиск на парцеллированную конструкцию в корпусе невозможно.

Объем собранного материала составил десятки тысяч страниц, что нереально обработать одному исследователю вручную за ограниченный период времени. Неизбежно возникла задача автоматизации процесса его обработки, другими словами, создания программного инструмента, способного справиться с поставленными задачами. Решением проблемы создания сопоставительного корпуса стала программа *Linguistica* (Свидетельство о государственной регистрации № 2014660349 от 06.10.2014). Идея, разработка концепции программы, интерфейса, тестирование и т. д., а также правообладание: Ю. В. Богоявленская; автор-программист: С. А. Александров. Программа представляет собой современный технологический инструмент, предназначенный для создания сопоставительных лингвистических корпусов, их обработки и получения статистических данных. Программа позволяет исследователю самостоятельно выстраивать деревья зависимостей (параметрические деревья) и снабжать разметкой как интересующие фрагменты, так и тексты. В программе имеется система поиска, статистическая обработка результатов; есть возможность получения конкорданса – генерируемого программой списка фрагментов по заданным параметрам с доступом к источнику. Охарактеризуем созданный при помощи программы корпус.

Таблица 1 – Типологические характеристики корпуса, сформированного при помощи программы Linguistica

Сопоставительный	Корпус включает два подкорпуса текстов на французском и русском языках, предназначенных для сопоставления.
Специализированный	В корпусе представлены тексты газетного дискурса.
Открытый	Корпус постоянно пополняется новыми текстами.
Динамический	Корпус предназначен для изучения синхронной динамики парцелляции в газетных текстах.
Полнотекстовый	Корпус включает полные тексты французских и русских газетных статей.
Исследовательский	Корпус представляет собой богатую эмпирическую базу, достаточную для исследования феномена парцелляции в двух языках.
Корпус ручной сборки	Отбор текстов производится в неавтоматическом режиме. Тексты размещаются в электронной коллекции в формате word, затем заносятся в программу. Возможно копирование текста с сайта прямо в программу.
Корпус ручной разметки	Разметка осуществляется вручную при помощи специальных инструментов. Виды используемой корпусной разметки: экстралингвистическая и лингвистическая.

Корпус описывается при помощи эмпирически наблюдаемых и количественно измеримых параметров: хронологического, гендерного, конструктивного, семантического, риторического, дискурсивного и др.

Программа создавалась специально для решения задач данного исследования, посвященного сопоставительному анализу парцелляции для выяснения того, как в действительности используется парцелляция носителями языка в речевой деятельности. Однако ее потенциал оказался намного шире. Она была протестирована и показала себя не менее эффективной в семантико-когнитивных и синтаксических исследованиях: прецедентных концептов [26], прецедентных высказываний [27, 28] и абсолютных конструкций с причастием [29], для которых создавались французский, русский и английский корпусы.

Для более точной характеристики корпуса мы вводим понятие *«объектно-ориентированный корпус»*, под которым понимаем *специализированный корпус, предназначенный для изучения конкретного лингвистического объекта.* Такой корпус является объектом, выступающим в качестве модели некоторой внешней по отношению к нему лингвистической реальности. В силу направленности на сопоставительное исследование объекта, корпус представляет собой модель двух языковых пространств. Следовательно, созданный корпус является *сопоставительным объектно-ориентированным корпусом (СООК)*.

3. Принципы формирования СООК

Как известно, любой корпус должен отражать речевую деятельность человека – сложный семиотический объект. Решение этой задачи, а также обеспечение достоверности полученных на материале корпуса данных определяется в научной литературе такими критериями, как *репрезентативность и сбалансированность*. В ходе создания собственного сопоставительного корпуса, включающего два подкорпуса, мы пришли к выводу о необходимости учета не только двух вышеперечисленных, но и *принципа сопоставимости корпусов*. С целью оценки достоверности и репрезентативности полученных результатов была разработана специальная *методика верификации достоверности полученных данных*, которая также позволяет регулировать объем корпуса [30]. Рассмотрим выше обозначенные принципы и технологию их применения в нашем корпусе.

Отметим, что проблема репрезентативности и сбалансированности корпусов находится в зоне повышенного внимания исследователей с 1990-х гг. Обобщая имеющиеся в лингвистической литературе дефиниции, определим *репрезентативность* как *возможность*

распространения представления о части (*корпусе*) *на целое* (*язык или его часть*). Другими словами, выборка должна полно и достоверно отображать признаки той совокупности, частью которой она является. Для специализированных корпусов это требование может звучать как *необходимо-достаточное количество текстов, обеспечивающих решение исследовательских задач*.

В ряде работ, посвященных данной проблеме, отмечается, что *в целом апробированных способов обеспечения репрезентативности корпусов не предложено* [31, 32]. Особенно подчеркивается, что применительно к общеязыковому (универсальному) корпусу это понятие невозможно рассчитать и описать строго математически. Признать корпус репрезентативным можно только в том случае, если полученные на материале корпуса статистические данные объективно отражают лингвистическую реальность. Другими словами, опираясь на эти данные, можно сделать обоснованные выводы об изучаемом феномене в целом и его закономерностях. Для проверки достоверности статических данных и оценки репрезентативности корпуса мы предлагаем использовать разработанную нами специальную методику верификации достоверности полученных при индексации корпуса данных, которая подробно описана в нашей статье [30].

Переходим к обсуждению *принципа сбалансированности*. Под сбалансированностью понимается *пропорциональное представление в корпусе текстов различных периодов, жанров, стилей, авторов* и т. п. Достижение этого параметра предполагается за счет включения текстов, имеющих различную жанровую принадлежность. Естественно, что при решении такой задачи остро встает проблема национальных систем жанров, приведения их к единой классификации, которая должна лечь в основу построения жанрового параметрического дерева для индексирования в корпусе. Проблема осложняется вопросом соответствия / несоответствия жанровых характеристик, количественным составом жанров в национальной газетной прессе.

Ряд исследователей обращают особое внимание на необходимость пропорционального включения текстов разных жанров в состав корпуса для достижения сбалансированности. На наш взгляд, наилучшим решением является *сплошная выборка текстов, содержащих исследуемый объект*, из каждого анализируемого газетного номера. Такой подход позволяет выявить реальную картину функционирования объекта (в нашем случае парцелляции) в газетных текстах различных жанров. Представляется, что искусственно определенное пропорциональное жанровое наполнение корпуса может привести к недостоверным результатам, хотя понимание сбалансированности корпуса во многом определяется задачами, которые ставит перед собой исследователь.

При обеспечении репрезентативности и сбалансированности текстов в структуре корпуса важными являются также *хронологические ограничения*, поскольку корпус должен охватывать и характеризовать какой-либо определенный период времени. Ввиду стремления отразить динамику развития парцелляции на синхроническом срезе, мы определили временную принадлежность текстов 20-летним периодом, с 1995 по 2015 гг. Это поставило перед создателем корпуса дополнительную проблему сбора материала, так как не все газеты имеют полнообъемные, глубокие и доступные архивные коллекции, что повлияло на сужение круга источников. Для исследования был отобран ряд респектабельных и массовых газет, имеющих архивы соответствующей глубины.

Переходим к третьему принципу. В зарубежной научной литературе обсуждаются также принципы сопоставимости корпусов, позволяющие добиться достаточной однородности корпусного текстового массива и, соответственно, обеспечивающие точность и адекватность выводов относительно сопоставляемых феноменов. А. МакЭнери выдвигает следующие критерии сопоставимости: 1) *корпусы должны содержать компоненты, отобранные по одному принципу;* 2) *иметь аналогичную сбалансированность и репрезентативность, т. е. одинаковые пропорции текстов одних и тех же жанров, отобранные в определенный период и* 3) *не должны быть переводными* [33, с. 450]. С точки зрения Б. Абера, построить репрезентативный сопоставительный корпус – это, прежде всего, определиться с границами

бытования языкового факта, изучению которого посвящается исследование. Необходимо ограничить область, тематику, жанр текстов и т. д., которые будут включены в корпус, что позволит обеспечить лингвистическую гомогенность корпуса [13, с. 21]. Таким образом, критерий сопоставимости сводится к *тематической и стилистической однородности текстов*. Э. Дежан, Э. Госсье предлагают в качестве критерия сопоставимости использовать то, что авторы назвали *критерием лексической близости*: корпусы должны содержать в значительной степени близкую лексику [25]. Английский корпусный лингвист А. Килгаррифф предлагает ввести *принцип симилярности (подобия) сопоставительных корпусов* и ищет способы ее исчисления. Автор приходит к выводу, что какова бы ни была методика ее исчисления, суждение о симилярности корпусов будет иметь субъективный характер, т.к. какие-то элементы под одним углом рассмотрения будут выглядеть подобными, но не поддающимися сравнению под другим [14, с. 233, 248–249].

Как показывают проанализированные работы, определение сопоставимости корпусов проистекает из целей их создания. Какими бы ни были тип и содержание корпуса, самым главным для исследователя должна быть конечная цель, которая и определяет характеристики создаваемого корпуса. Для нашего исследования тематический критерий, выдвинутый в [13, 33], значительно сузил бы материал исследования и не позволил бы рассмотреть функционирование парцелляции в различных жанрах. Близость лексики – критерий, выдвигаемый в [25, 14], – не является для нас решающим фактором, поскольку исследование носит не лексико-семантический характер, а ориентировано на изучение парцелляции как когнитивно-дискурсивного феномена.

Итак, сопоставимость – важный принцип, которого необходимо придерживаться при формировании корпусов данного типа. Учитывая вышесказанное, мы реализовали принцип сопоставимости корпусов, руководствуясь следующими критериями:

1) *квалитативные критерии*:

– отбор материала для сопоставляемых корпусов должен осуществляться по единому принципу: в корпус включаются только тексты, содержащие хотя бы одну парцелляцию;

– материал для сопоставления отбирается из соотносимых источников: респектабельных и массовых общественно-политических газет;

– источники не должны содержать переводных статей.

2) *квантитативный критерий:*

– репрезентативный (не обязательно равный) объем корпусов, регулируемый специальной методикой верификации полученных в результате корпусного анализа данных.

Заключение

Как мы убедились, имеющиеся и доступные в сети Интернет корпусы различаются по своим типологическим характеристикам, содержанию, структуре, объему, хронологическим рамкам и т. д., что существенно осложняет или делает невозможным их использование в сопоставительных корпусных исследованиях. Для такого вида исследования необходимо создание собственного корпуса, отвечающего на потребности и задачи, которые ставит перед собой лингвист. Решение выявленной проблемы становится возможным при использовании программы *Linguistica*, предназначенной для создания сопоставительных лингвистических корпусов, их обработки и получения статистических данных. Программа позволяет исследователю самостоятельно снабжать разметкой по предварительно выстроенным параметрическим деревьям как интересующие элементы, фрагменты, так и тексты в целом. Удобная система поиска, статистической обработки результатов, возможность получения конкорданса, т. е. генерируемого программой списка элементов / фрагментов / текстов по заданным параметрам, доступ к источнику прямо из конкорданса и многое другое делает из данной программы простой и эффективный инструмент работы с СООК. Созданный при помощи программы СООК обладает следующими типологическими характеристиками: открытый, динамический, полнотекстовый, исследовательский, корпус ручной сборки и ручной разметки.

В предложенной типологии многоязычных корпусов сопоставительный корпус (в т. ч. СООК) противостоит параллельным (выровненным) и комбинированным корпусам, используемым преимущественно для исследований различных аспектов перевода и для обучения методам и приемам перевода. Создание сопоставительного корпуса должно опираться не только на принципы репрезентативности и сбалансированности, но и на принцип сопоставимости (квантитативные и квалитативные критерии). Последовательная реализация данных принципов позволит обеспечить достоверность получаемых в ходе корпусного анализа данных об исследуемом объекте.

Дальнейшие перспективы предпринятого исследования мы видим в усовершенствовании интерфейса программы, встраивании расчета индекса достоверности, расширении инструментария ввода и обработки данных, включении полезных и удобных для исследователя инструментов и функций.

Литература:

1. Fludernik, M. *The Fictions of Language and the Languages of Fiction*. London, N.Y.: Routledge, 1993, 431 p.
2. Французский корпус Лейпцигского университета [Электронный ресурс]. Режим доступа: http://wortschatz.uni-leipzig.de/ws_fra/.
3. Корпус Frantext [Электронный ресурс]. Режим доступа: http://www.frantext.fr.
4. Корпус французского языка Lexiqum [Электронный ресурс]. Режим доступа: http://retour.iro.umontreal.ca/cgi-bin/lexiqum.
5. Британский Национальный корпус [Электронный ресурс]. Режим доступа: http://www.natcorp.ox.ac.uk/.
6. Корпус современного американского английского языка COCA [Электронный ресурс]. Режим доступа: http://corpus.byu.edu/coca/.
7. Национальный корпус русского языка [Электронный ресурс]. Режим доступа: http://www.ruscorpora.ru/.
8. Компьютерный корпус газетных текстов русского языка конца XX века [Электронный ресурс]. Режим доступа: http://www.philol.msu.ru/~lex/corpus/.
9. Gœuriot, L. *Découverte et caractérisation des corpus comparables spécialisés*: thèse de Doctorat. Université de Nantes, 2009, 152 p., www.tel.archives-ouvertes.fr/docs/00/47/ 44/05/PDF/these-lorraine-goeuriot.pdf.
10. Afli, H., Barrault, L., Schwenk, H. “Traduction automatique à partir de corpus comparables: extraction de phrases parallèles à partir de données comparables multimodales. ” *Actes de la conférence conjointe JEP-TALN-RECITAL*, Grenoble, 4–8 juin 2012, vol. 2,TALN, 2012, pp. 447–454.
11. Cartoni, B., Deléger, L. “Découverte de patrons paraphrastiques en corpus comparable: une approche basée sur les n-grammes.” *TALN*, 27 juin – 1er juillet 2011, Montpellier, 2011, http://www.atala.org/taln_archives/TALN/TALN-2011/taln-2011-court-031.pdf.
12. McEnery, A., Xiao, Z. “Parallel and Comparable Corpora: What is Happening.” *Incorporating Corpora: Translation and the Linguist*. Clevedon: Multilingual Matters, 2007, pp. 18–31, http://someya-net.com/104IT_Kansai_Initiative/corpora_ and_translation.pdf.
13. Habert, B. *Linguistique sur corpus. Études et réflexions*. Perpignan: Presses Universitaires de Perpignan, 2000, pp. 11–58.
14. Kilgariff, A. “Comparing Corpora.” *International Journal of Corpus Linguistics*. 2001, no. 6, pp. 97–133.
15. Bowker, L., Pearson, J. *Working with Specialized Language: A Practical Guide to Using Corpora*. London, N.Y.: Routeledge, 2002, 242 p.
16. Peters, C., Picchi, E., Biagini, L. “Parallel and Comparable Bilingual Corpora in Language Teaching and Learning.” *Proceedings of Teaching and Language Corpora*, 1996, pp. 68–82.
17. Rauf, S. *Efficient Corpus Selection for Statistical Machine Translation*: thèse de Doctorat. Université du Maine, 2012, https://tel.archives-ouvertes.fr/tel-00732984/.
18. Barzilay, R., Lee, L. *Learning to Paraphrase: An Unsupervised Approach Using Multiple-Sequence Alignment*. Edmonton, 2003, pp. 16–23.
19. Elhadad N., Sutaria K. “Mining a Lexicon of Technical Terms and Lay Equivalents.” *ACL BioNLP Workshop*. Prague, 2007, pp. 49–56.

20. Zanettin, F. “Bilingual Corpora and the Training of Translators.” *Meta*, vol. 4, no. 43, 1998, pp. 616–630.
21. Culo, O., Schirra, S.H., Neumann, S., Vela, M. “Empirical Studies on Language Contrast Using the English-German Comparable and Parallel Corpus.” *Proceedings of the LREC Workshop on Comparable Corpora*, 2008, pp. 47–51.
22. Guidère, M. *Introduction à la traductologie. Penser la traduction: hier, aujourd’hui, demain.* De Boeck Université, 2010.176 p.
23. Scarpa, F. *La traduction spécialisée. Une approche professionnelle à l’enseignement de la traduction*. Traduit et adapté par M.A. Fiola. Ottawa: Presses de l’Université d’Ottawa, 2010, www.books.google.fr.
24. Fung, P. “A Statistical View on Bilingual Lexicon Extraction.” *Parallel Corpora to Non-Parallel Corpora: Conference of the Association of Translation in the Americas (AMTA)*, 1998, pp. 1–17.
25. Déjean, H., Gaussier, E. “Une nouvelle approche à l’extraction de lexiques bilingues à partir de corpus comparables.” *Lexicometrica*, 2002, http://lexicometrica.univ-paris3.fr/thema/thema6/Dejean.pdf.
26. Богоявленская Ю. В., Буженинов А. Э. Прецедентное имя «Наполеон» в исторической памяти Франции: Опыт корпусного исследования // Политическая лингвистика. 2015. N 2. С. 137–143.
27. Богоявленская Ю. В., Чудинов А. П. Функционирование прецедентных высказываний во французском медиадискурсе // Вестник Нижегородского лингвистического университета им. Н. А. Добролюбова. 2015. Вып. 29. С. 20–35.
28. Bogoyavlenskaya, Yu. V., Nakhimova, E. A., Chudinov, A. P. Precedent utterances in the national historical memory: a corpus study // Вопросы когнитивной лингвистики. 2016. N 2. С. 39–48.
29. Нелюбина М. С., Богоявленская Ю. В. Абсолютная причастная конструкция и смежные явления во французском языке // Вестник Томского государственного университета. Филология. 2016. N 3(41). С. 27–36.
30. Богоявленская Ю.В. Репрезентативность лингвистического корпуса: метод верификации достоверности полученных данных // Политическая лингвистика. 2016. N 4. С. 163–166.
31. Беликов В. И., Копылов Н. Ю., Пиперски А. Ч., Селегей В. П., Шаров С. А. Корпус как язык: от масштабируемости к дифференциальной полноте // Компьютерная лингвистика и интеллектуальные технологии. Т. 1. N 12(19). 2013. С. 84–95.
32. Arbach, N., Ali, S. “Aspects théoriques et méthodologiques de la représentativité des corpus CORELA.” *Statut et utilisation des corpus en linguistique*, 2014, http://corela.revues.org/3029?lang=en.
33. McEnery, A. M. “Corpus Linguistics.” *The Oxford Handbook of Computational Linguistics*. Oxford: Oxford University Press, 2003, pp. 448–463.

Bogoyavlenskaya Yu.V.
Ural State Pedagogical University, Yekaterinburg, Russia

COMPARATIVE OBJECT-ORIENTED CORPUS: DEFINITION OF THE CONCEPT AND PRINCIPLES OF FORMATION

The article deals with the problems of access to corpora, published in free access for researches as part of comparative corpus linguistics and proposes their solution. The corpora of different languages represented on the Internet differ significantly in the goals and principles of their creation, volume, structure, content, types of marking, chronological framework and other characteristics that cannot ensure the objectivity of the data obtained through their analysis in comparative researches. At the first stage of the study, we analyze the works devoted to the principles of organization and content of a comparative corpus, we determine this concept and its place in the typology of multilingual corpora. In the next stage, setting the goal of creating our own comparative corpus to study parceling as a cognitive-discursive phenomenon, we introduce the concept of a comparative object-oriented corpus, meaning a specialized bilingual or multilingual corpus, devoted to the study of a specific linguistic object. Further, the principles of creating a corpus with account of its specificity are formulated: the principles of representativeness, symmetry and comparability. The last principle is based on the observance of a number of quantitative and

qualitative criteria for assembling this type of corpus. Only a consistent implementation of these principles will ensure the reliability of the data obtained in the course of the corpus analysis of the object under study. To build a comparative object-oriented corpus and its processing, we use the computer program *Linguistica* specially elaborated for solving research problems.

Key words: (comparative) corpus linguistics, comparative corpus, comparative object-oriented corpus, principles of corpus building.

About the author: **Bogoyavlenskaya Yuliya Valer'evna**, Doctor of Philology, Associate Professor of the Department of Romance Languages, Ural State Pedagogical University (Yekaterinburg, Russia); e-mail: jvbog@yandex.ru, ORCID ID 0000-0002-4500-1491.

References:

1. Fludernik, M. *The Fictions of Language and the Languages of Fiction*. London, N.Y.: Routledge, 1993, 431 p.
2. Corpus of French of University of Leipzig, http://wortschatz.uni-leipzig.de/ws_fra/.
3. Corpus Frantext, http://www.frantext.fr.
4. Corpus of French Lexiqum, http://retour.iro.umontreal.ca/cgi-bin/lexiqum.
5. British National Corpus, http://www.natcorp.ox.ac.uk/.
6. Corpus of Contemporary American English (COCA), http://corpus.byu.edu/coca/.
7. Russian National Corpus, http://www.ruscorpora.ru/.
8. Computer Corpus of Newspaper Texts of the Russian Language of the End of XX Century, http://www.philol.msu.ru/~lex/corpus/.
9. Gœuriot, L. *Découverte et caractérisation des corpus comparables spécialisés*: thèse de Doctorat. Université de Nantes, 2009, 152 p., www.tel.archives-ouvertes.fr/docs/00/47/ 44/05/PDF/these-lorraine-goeuriot.pdf.
10. Afli, H., Barrault, L., Schwenk, H. "Traduction utomatique à partir de corpus comparables: extraction de phrases parallèles à partir de données comparables multimodales. " *Actes de la conférence conjointe JEP-TALN-RECITAL*, Grenoble, 4-8 juin 2012, vol. 2,TALN, 2012, pp. 447–454.
11. Cartoni, B., Deléger, L. "Découverte de patrons paraphrastiques en corpus comparable: une approche basée sur les n-grammes." *TALN*, 27 juin – 1er juillet 2011, Montpellier, 2011, http://www.atala.org/taln_archives/TALN/TALN-2011/taln-2011-court-031.pdf.
12. McEnery, A., Xiao, Z. "Parallel and Comparable Corpora: What is Happening." *Incorporating Corpora: Translation and the Linguist*. Clevedon: Multilingual Matters, 2007, pp. 18–31, http://someya-net.com/104IT_Kansai_Initiative/corpora_ and_translation.pdf.
13. Habert, B. *Linguistique sur corpus. Études et réflexions*. Perpignan: Presses Universitaires de Perpignan, 2000, pp. 11–58.
14. Kilgariff, A. "Comparing Corpora." *International Journal of Corpus Linguistics*. 2001, no. 6, pp. 97–133.
15. Bowker, L., Pearson, J. *Working with Specialized Language: A Practical Guide to Using Corpora*. London, N.Y.: Routeledge, 2002, 242 p.
16. Peters, C., Picchi, E., Biagini, L. "Parallel and Comparable Bilingual Corpora in Language Teaching and Learning." *Proceedings of Teaching and Language Corpora*, 1996, pp. 68–82.
17. Rauf, S. *Efficient Corpus Selection for Statistical Machine Translation*: thèse de Doctorat. Université du Maine, 2012, https://tel.archives-ouvertes.fr/tel-00732984/.
18. Barzilay, R., Lee, L. *Learning to Paraphrase: An Unsupervised Approach Using Multiple-Sequence Alignment*. Edmonton, 2003, pp. 16–23.
19. Elhadad N., Sutaria K. "Mining a Lexicon of Technical Terms and Lay Equivalents." *ACL BioNLP Workshop*. Prague, 2007, pp. 49–56.
20. Zanettin, F. "Bilingual Corpora and the Training of Translators." *Meta*, vol. 4, no. 43, 1998, pp. 616–630.
21. Culo, O., Schirra, S.H., Neumann, S., Vela, M. "Empirical Studies on Language Contrast Using the English-German Comparable and Parallel Corpus." *Proceedings of the LREC Workshop on Comparable Corpora*, 2008, pp. 47–51.
22. Guidère, M. *Introduction à la traductologie. Penser la traduction: hier, aujourd'hui, demain.* De Boeck Université, 2010. 176 p.

23. Scarpa, F. *La traduction spécialisée. Une approche professionnelle à l'enseignement de la traduction*. Traduit et adapté par M.A. Fiola. Ottawa: Presses de l'Université d'Ottawa, 2010, www.books.google.fr.
24. Fung, P. "A Statistical View on Bilingual Lexicon Extraction." *Parallel Corpora to Non-Parallel Corpora: Conference of the Association of Translation in the Americas (AMTA)*, 1998, pp. 1–17.
25. Déjean, H., Gaussier, E. "Une nouvelle approche à l'extraction de lexiques bilingues à partir de corpus comparables." *Lexicometrica*, 2002, http://lexicometrica.univ-paris3.fr/thema/thema6/Dejean.pdf.
26. Bogoyavlenskaya, Yu. V., Buwheninov A. E. "Precedent Name "Napoleon" in the Historical Memory of France: Experience of Corpus-Based Research." *Politicheskaja lingvistika*, no. 2, 2015, pp. 137–143.
27. Bogoyavlenskaya, Yu. V., Chudinov, A. P. "Functionning of Precedential Statements in French Media Discourse." *Vestnik Nizhegorodskogo Lingvisticheskogo Universiteta Im. N. A. Dobroljubova*, no. 29, 2015, pp. 20–35.
28. Bogoyavlenskaya, Yu. V., Nakhimova, E. A., Chudinov, A. P. "Precedent Utterances in the National Historical Memory: a Corpus Study. *Voprosy Kognitivnoj Lingvistiki*, no. 2, 2016, pp. 39–48.
29. Nelubina M. S., Bogoyavlenskaya, Yu. V. "The Absolute Participle Construction in the French Language and the Adjoining Phenomena." *Vestnik Tomskogo Gosudarstvennogo Universiteta. Filologija*, vol. 3, no. 41, 2016, pp. 27–36.
30. Bogoyavlenskaya, Yu. V. "Representativeness of Text Corpus: Method of Verification of Reliability of Data." *Politicheskaja lingvistika*, no. 4, 2016, pp. 163–166.
31. Belikov, V., Kopylov, N., Piperski, A., Selegey, V., Sharoff, S. "Corpus as Language: from Scalability to Register Variation." *Komp'juternaja lingvistika i intellektual'nye tehnologii*, Issue 12, vol. 1, no. 19, 2013, pp. 84–95.
32. Arbach, N., Ali, S. "Aspects théoriques et méthodologiques de la représentativité des corpus CORELA." *Statut et utilisation des corpus en linguistique*, 2014, http://corela.revues.org/3029?lang=en.
33. McEnery, A. M. "Corpus Linguistics." *The Oxford Handbook of Computational Linguistics*. Oxford: Oxford University Press, 2003, pp. 448–463.